Exact enumeration of cherries and pitchforks in ranked trees under the coalescent model

Filippo Disanto* Thomas Wiehe*

August 21, 2012

Abstract

We consider exact enumerations and probabilistic properties of *ranked* trees when generated under the random coalescent process. Using a new approach (see [9, 10]), based on generating functions, we derive several statistics such as the exact probability of finding k *cherries* in a ranked tree of fixed size n. We then extend our method to consider also the number of *pitchforks*. We find a recursive formula to calculate the joint and conditional probabilities of cherries and pitchforks when the size of the tree is fixed.

1 Introduction

Given a direction by time, ancestry relationship between species, individuals, alleles or cells can be depicted as a rooted tree. Of particular interest are binary rooted unordered trees. These can be further classified into several subclasses. Here we will *ranked trees*, which are defined below.

We assume that trees are generated by the coalescent process.

An important parameter is the number of *cherries* of a tree. By a new approach based on generating functions we extend previous results (see for example [9]) deriving an exact formula for the probability of finding k cherries in a ranked tree of size n. Furthermore, we show that several known statistics (see [10]) concerning *pitchforks* follow as corollaries from a partial differential equation which also gives an efficient recursion to compute the conditional probability distribution of pitchforks given a certain number of cherries.

*Institut für Genetik, Universität zu Köln; Zülpicher Straße 47a, 50674 Köln, Germany

One motivation for this study comes from population genetics and the question how 'typical' *coalescent trees* [13] look like. Our results give some insight into structural properties of trees generated under the standard neutral model [12]. These results provide a reference against which non neutral and/or non independently generated trees may be compared. To illustrate the latter we pay attention to trees which are linked along a recombining chromosome.

2 Preliminaries

We start with some basic definitions. A *binary rooted* tree is a tree with a root and in which all nodes have outdegree either 0 or 2. Nodes with outdegree 2 are called *internal*, nodes with outdegree 0 are *external*. External nodes are also called *leaves*. The size n of a tree is the number of its external nodes. The *subtree* of an internal node i is the tree with root i. A tree is said to be *un-ordered* when it is taken in the graph theoretic sense so that subtrees stemming from an internal node have not a left-right order between themselves. Here, we care about tree topology and we do not care about branch lengths. We consider the following class. A binary un-ordered tree of size n is said to be a *ranked tree* if the set of internal nodes is totally ordered by labels belonging to $\{1, 2, ..., n\}$ in such a way that each child's label is greater than its parent's label, (see Fig. 1). The total order of internal labels can be interpreted as a historical time order; accordingly, Harding [4] calls such trees *histories*.

We will denote by $\mathcal{R}$ the set of ranked trees and by $\mathcal{R}_n$ the set of trees of size n. In what follows, $n = n(t)$ always represents the number of leaves of a ranked tree t.

The cardinality of the set $\mathcal{R}_n$ is given by the following exponential generating function

$$\mathcal{R}(x) = \sum_{n \geq 0} \frac{|\mathcal{R}_n|}{n!} x^n = \sec(x) + \tan(x). \tag{1}$$

whose first coefficients $|\mathcal{R}_n|$ (with $n > 0$) are

$$1, 1, 1, 2, 5, 16, 61, 272, ...$$

Ranked trees can be bijectively mapped to 0-1-2-*increasing trees* (see Callan, 2005; `http://www.stat.wisc.edu/~callan/notes`). From this, it follows that the numbers given by (1) correspond to sequence $A000111$ in Sloane [11] and are known as *Euler numbers*.

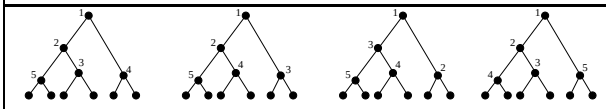
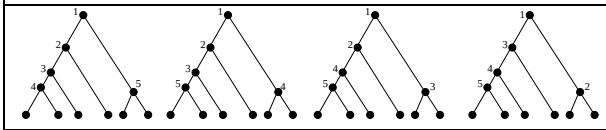
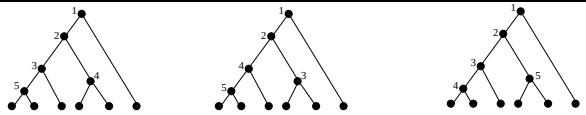
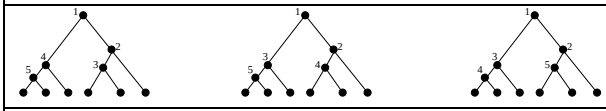
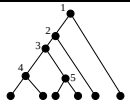
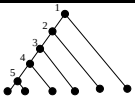
trees	# cherries	# pitchforks
	3	0
	2	1
	2	1
	2	2
	2	0
	1	1

Figure 1: The sixteen possible ranked trees of size six classified by shape. Within each class all possible orderings of the internal nodes are displayed. Number of cherries and pitchforks are indicated.

2.1 Trees as a result of the coalescent process

The coalescent of size n is a model for the genealogical history of a sample of n genes. It has been introduced in population genetics by Kingman and Ewens [7, 8] and has nowadays textbook status [13]. Ranked trees can be generated by the *coalescent process*, which starts with n leaves and works by successively coalescing two randomly chosen branches until it reaches the 'most recent common ancestor' when the last two remaining branches are joined.

To reflect time order one can assign an integer to each internal node when created, for instance the label $n-1$ to the first coalescent event and 1 to the last event, the most recent common ancestor, or the root of the tree.

The probability distribution of ranked trees $P_{\mathcal{R}}$ generated under the coalescent process is essentially contained in the paper of Tajima [12] and it is described below.

Probability distribution of ranked trees

Let $t \in \mathcal{R}$ and let $o(t)$ be the number of internal nodes i whose children are two leaves. Such internal nodes are called the *cherries* of the tree. For example, (see Fig. 1). Given $t \in \mathcal{R}_n$, from Tajima [12] follows that

$$P_{\mathcal{R}}(t) = \frac{2^{n-1-o(t)}}{(n-1)!}\,, \tag{2}$$

i.e. the probability of any ranked tree $t \in \mathcal{R}_n$ depends only on two parameters, o and n.

The probability of generating the same ranked trees twice

Considering trees linked on a common chromosome one observes that chromosomal linkage substantially increases the probability that two 'neighboring' trees are identical even if separated by a recombination event. To quantify the effect of linkage and recombination it is important to know the background probability that two independently generated trees are identical. This probability can be found with the help of the genarating function

$$Y(x,z) = \sum_{t\in\mathcal{R}} \frac{x^{o(t)} z^{n(t)-1}}{(n(t)-1)!},$$

discussed in more details in Section 3.1.1, eq. (6).

We have the following result.

Proposition 1 *The probability that two independently generated ranked trees of size n are identical is*

$$p_n = \frac{4^{n-1}}{(n-1)!} \times [z^{n-1}]Y\left(\frac{1}{4}, z\right) .$$

Proof: From eq. (2) the probability that $t_1, t_2 \in \mathcal{R}_n$ are identical is

$$\begin{aligned} p_n &= \sum_{t\in\mathcal{R}_n} P_{\mathcal{R}}(t)^2 \\ &= \frac{1}{(n-1)!^2} \sum_{t\in\mathcal{R}_n} 4^{n-1-o(t)} \\ &= \frac{4^{n-1}}{(n-1)!^2} \sum_{t\in\mathcal{R}_n} \left(\frac{1}{4}\right)^{o(t)} \\ &= \frac{4^{n-1}}{(n-1)!} \times [z^{n-1}]Y\left(\frac{1}{4}, z\right), \end{aligned}$$

where $[z^{n-1}]Y(1/4, z)$ means the $(n-1)$-st coefficient of the Taylor expansion of $Y(1/4, z)$ in $z = 0$. □

3 Enumerative results

3.1 Outdegree of the nodes in ranked and 0-1-2-increasing trees

Let $t \in \mathcal{R}_n$ and $m = n - 1$. Remove all leaves and external branches from t and obtain a reduced tree $\rho(t)$. The tree $\rho(t)$ is a so-called 0-1-2-*increasing tree* of size m, where, this time, the size is the total number of nodes in the tree and not only of the leaves. The class $\mathcal{I}_{012}$ of 0-1-2-increasing trees is composed of un-ordered rooted trees where all nodes have outdegree 0, 1 or 2. The m nodes of such a tree carry totally ordered labels belonging to $\{1, 2, ..., m\}$. Moreover, the labelling is such that any child node label is greater than that of the parent node. As usual $\mathcal{I}_{012_m}$ denotes the set of 0-1-2-increasing trees of size m. Hence, the function ρ is a bijection from $\mathcal{R}_n$ to $\mathcal{I}_{012_m}$.

Given a ranked tree t, the *outdegree of an internal node* of t is the outdegree of the corresponding node in $\rho(t)$. Thus, if $t \in \mathcal{R}$, the nodes of outdegree 0 (resp. 1, 2) are defined as the nodes with 2 (resp. 1, 0) leaves as direct descendants.

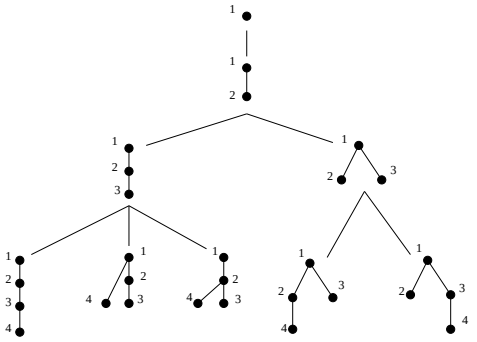


Figure 2: First levels of the generating tree associated to Θ.

Here, we derive the enumeration of 0-1-2-increasing trees with respect to the size and to the number of nodes with outdegree 0,1 and 2. The bijection ρ will allow us to use this enumerative result in Section 3.1.2 to determine the probability distribution of the random variable o, the *number of cherries*, when t is a ranked tree of size n generated by the coalescent process. It is already known (see McKenzie [9]) that $o(t)$ is asymptotically normal for large n.

3.1.1 Recursive construction of 0-1-2-increasing trees

We show now how the class of 0-1-2-increasing trees can be generated recursively. In particular we construct each tree belonging to $\mathcal{I}_{012_{m+1}}$ by adding a new node to some tree in $\mathcal{I}_{012_m}$. This construction, denoted by Θ, will then be translated into a functional equation. Solving the equation we obtain a bivariate exponential generating function counting the considered increasing trees with respect to size and to the number of nodes with outdegree 0, 1 and 2.

Given a tree $t \in \mathcal{I}_{012_m}$, Θ simply adds the node labelled '$m+1$' as a child of a node of t having outdegree less than two. Let $o(t), p(t)$ and $q(t)$ denote the number of nodes with outdegree $0, 1$ and 2 respectively. Θ applied to t produces $o(t) + p(t)$ elements of $\mathcal{I}_{012_{m+1}}$ each time adding the new node labelled $m + 1$ as a child of the nodes counted in $o(t) + p(t)$. In Fig. 2 we depict the first steps of this construction process.

Note that $o(t) = q(t)+1$ and $o(t)+p(t)+q(t) = m$. From these relations we have, in particular, that $p(t) = m-2o(t)+1$. The construction Θ can be translated into the following *succession rule* (see Banderier et al. [1]) where each tree is represented by a label composed of the values of its parameters

o and m while the exponents show how many times the label is produced,

$$(o,m) \to (o,m+1)^o\,(o+1,m+1)^{m-2o+1}.$$

In particular, given a tree t with parameters $o = o(t)$ and $m = m(t)$, the application of Θ to t produces o new trees having size $m+1$ and o cherries and $m-2o+1$ new trees having size $m+1$ and $o+1$. The starting point of the construction is the unique tree of size one represented by $(1,1)$.

Now consider the exponential generating function

$$Y(x,z) = \sum_{t \in \mathcal{I}_{012}} \frac{x^{o(t)} z^{m(t)}}{m(t)\,!}.$$

The previous succession rule can be translated as follows into an equation for $Y(x,z)$.

$$\begin{aligned} Y(x,z) &= xz + \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{o x^o z^{m+1}}{(m+1)\,!} + \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{(m-2o+1)(x^{o+1} z^{m+1})}{(m+1)\,!} \\ &= xz + (1-2x) \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{o x^o z^{m+1}}{(m+1)\,!} + xz \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{x^o z^m}{m\,!} \\ &= xz + (1-2x) \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{o x^o z^{m+1}}{(m+1)\,!} + xz\,Y(x,z) \end{aligned}$$

From the previous equation we obtain that

$$\frac{Y(x,z)(1-xz) - xz}{1-2x} = \sum_{x^o z^m \in \mathcal{I}_{012}} \frac{o x^o z^{m+1}}{(m+1)\,!}.$$

Differentiating both sides with respect to the variable z we have

$$\frac{1}{1-2x}\left(\frac{dY}{dz}(x,z)\,(1-xz) - x\,Y(x,z) - x\right) = x\,\frac{dY}{dx}(x,z),$$

which is equivalent to

$$x(1-2x)\,\frac{dY}{dx}(x,z) + (xz-1)\,\frac{dY}{dz}(x,z) = -x\,Y(x,z) - x. \qquad (3)$$

The previous first order partial differential equation can be solved using the *method of characteristics* (see [2]) respecting the condition given by eq. (1)

$$Y(1,z) = \sec(z) + \tan(z) - 1.$$

Indeed $Y(1,z)$ must represent the exponential generating function counting 0-1-2-increasing trees with respect to size.

Applying the method consists, first, of solving the two following ordinary differential equations

$$\begin{aligned} z' &= \frac{xz-1}{x(1-2x)} \\ Y' &= \frac{-xY-x}{x(1-2x)} \end{aligned}$$

The solutions are

$$\begin{aligned} z &= \frac{c_1 + 2\,\arctan(\sqrt{2x-1})}{\sqrt{2x-1}}, \\ Y &= c_2\sqrt{2x-1} - 1, \end{aligned} \tag{4}$$

with constants c_1 and c_2 and where c_2 can be written as a function of c_1 in the following way

$$c_2 = G(c_1) = G(z\sqrt{2x-1} - 2\,\arctan(\sqrt{2x-1})).$$

In this way equation (4) becomes

$$Y(x,z) = G(z\sqrt{2x-1} - 2\,\arctan(\sqrt{2x-1}))\sqrt{2x-1} - 1,$$

which gives

$$\sec(z) + \tan(z) - 1 \;=\; Y(1,z) \;=\; G(z - \frac{\pi}{2}) - 1.$$

Function G must satisfy

$$G(z) = \sec(z + \frac{\pi}{2}) + \tan(z + \frac{\pi}{2}) = \frac{-1-\cos(z)}{\sin(z)}.$$

Inserting this into (4) we have

$$Y(x,z) = \sqrt{2x-1}\left(\frac{-1-\cos(z\sqrt{2x-1}-2\,\arctan(\sqrt{2x-1}))}{\sin(z\sqrt{2x-1}-2\,\arctan(\sqrt{2x-1}))}\right)-1,$$

which, after some calculations, finally gives

$$Y(x,z) = \frac{\sqrt{2x-1}}{\tan\left(-\frac{z\sqrt{2x-1}}{2}+\arctan(\sqrt{2x-1})\right)}-1. \tag{5}$$

Note that the condition $Y(1,z) = \sec(z)+\tan(z)-1$ is respected. Indeed

$$Y(1,z) = \frac{1}{\tan\left(-\frac{z}{2}+\frac{\pi}{4}\right)}-1$$

and

$$\begin{aligned}\frac{1}{\tan\left(-\frac{z}{2}+\frac{\pi}{4}\right)} &= \frac{1+\tan\left(\frac{z}{2}\right)}{1-\tan\left(\frac{z}{2}\right)} = \frac{1+\cos(z)+\sin(z)}{1+\cos(z)-\sin(z)}\\ &= \frac{1+\cos^2(z)+2\cos(z)-\sin^2(z)}{(1+\cos(z)-\sin(z))^2}\\ &= \frac{\cos(z)}{1-\sin(z)} = \frac{1+\sin(z)}{\cos(z)}\end{aligned}$$

Moreover, using the fact that

$$\exp\left(z\sqrt{-2x+1}\right) = \cos(z\sqrt{2x-1})+i\,\sin(z\sqrt{2x-1}),$$

we can write eq. (5) in terms of the exponential function as

$$Y(x,z) = \frac{2\left(x\exp\left(\sqrt{-2\,x+1}z\right)-x\right)}{\left(\sqrt{-2\,x+1}-1\right)\exp\left(\sqrt{-2\,x+1}z\right)+\sqrt{-2\,x+1}+1}. \tag{6}$$

Performing the substitution $x = 1/4$ we have that

$$Y\left(\frac{1}{4},z\right) = \frac{e^{\left(\sqrt{\frac{1}{2}}z\right)}-1}{2\left(\left(\sqrt{\frac{1}{2}}-1\right)e^{\left(\sqrt{\frac{1}{2}}z\right)}+\sqrt{\frac{1}{2}}+1\right)},$$

the Taylor expansion of which is

$$Y\left(\frac{1}{4}, z\right) = \frac{1}{4}z + \frac{1}{8}z^2 + \frac{5}{96}z^3 + \frac{1}{48}z^4 + \frac{1}{120}z^5 + \dots.$$

Using the result of Proposition 1 we can now effectively calculate the probability p_n that two ranked trees having n leaves are identical when generated independently by the coalescent process: $p_2 = \frac{4}{1!} \times \frac{1}{4} = 1, p_3 = \frac{4^2}{2!} \times \frac{1}{8} = 1, p_4 = \frac{4^3}{3!} \times \frac{5}{96} = \frac{5}{9}, p_5 = \frac{4^4}{4!} \times \frac{1}{48} = \frac{2}{9}$ and $p_6 = \frac{4^5}{5!} \times \frac{1}{120} = \frac{16}{225}$, and so on.

3.1.2 The probability distribution of the number of cherries

We are now ready to state the enumeration of ranked trees with respect to size and number of nodes of outdegree 0, 1 or 2, when each tree is weighted by its probability under the coalescent process. This exact enumerative result is novel and achieved with the help of the weighted generating function

$$F(x, z) = \sum_{t \in \mathcal{R}_n,\, n>1} \frac{2^{n(t)-1-o(t)}}{(n(t)-1)!} x^{o(t)} z^{n(t)}.$$

Function F has a more intuitive interpretation if one considers the transformation $Y_w = \frac{F}{z}$ instead. It can be interpreted as a weighted exponential generating function counting 0-1-2-increasing trees with respect to the outdegree and the total number of nodes.

Starting from equation (6), we perform some substitutions on Y to obtain Y_w. In particular we have $Y_w = Y\left(\frac{x}{2}, 2z\right)$ and, multiplying by z, we finally obtain the desired function F.

Proposition 2 *The weighted ordinary generating function of ranked trees considered with respect to size and number of cherries is*

$$F(x, z) = \frac{zx \exp\left(2z\sqrt{-x+1}\right) - zx}{\left(\sqrt{-x+1}-1\right)\exp\left(2z\sqrt{-x+1}\right) + 1 + \sqrt{-x+1}}. \qquad (7)$$

The probability of having o' cherries in a ranked tree of size n corresponds to the coefficient of $x^{o'}z^n$ in the Taylor expansion of F around $z = 0$, i.e.

$$P_n(o = o') = [x^{o'}z^n]F(x, z).$$

The first terms of the Taylor expansion of (7) are described below;

$$
\begin{aligned}
F(x,z) \;=\;& xz^2 \\
&+xz^3 \\
&+\frac{1}{3}\left(x^2+2\,x\right)z^4 \\
&+\frac{1}{3}\left(2\,x^2+x\right)z^5 \\
&+\frac{1}{15}\left(2\,x^3+11\,x^2+2\,x\right)z^6 \\
&+\frac{1}{45}\left(17\,x^3+26\,x^2+2\,x\right)z^7 \\
&+\frac{1}{315}\left(17\,x^4+180\,x^3+114\,x^2+4\,x\right)z^8 \\
&+\ldots\,.
\end{aligned}
$$

Looking at Fig. 1 one can check that, for example, there are exactly 11 trees represented by the monomial x^2z^6. Each one of them has probability $\frac{1}{15}$. This is in agreement with the term $\frac{11}{15}x^2z^6$ in the expansion. Indeed, $\frac{11}{15}$ is the probability to obtain a ranked tree of size 6 with two cherries.

Using the result of Proposition 2 we compute the discrete probability distribution of the random variable $o(t)$ for trees of fixed size n. In this case o is a random variable which takes values between 1 and $\lfloor n/2 \rfloor$. In Fig. 3 we have depicted the distribution of o for a ranked tree of size $n = 54$.

By Proposition 2 one can also determine the expected value $E_o(n)$ and the variance $Var_o(n)$ of the random variable o in dependence of tree size n. Using other methods these have been determined before, for example by McKenzie [9].

Using our approach the expectation is

$$E_o(n) = [z^n]\frac{dF}{dx}(1,z) = [z^n]\frac{z^4-3z^3+3z^2}{3(z-1)^2}\,.$$

If $n > 2$, this simplifies to

$$E_o(n) = \frac{n}{3}\,.$$

The second moment is

$$\begin{aligned} E_{o^2}(n) &= [z^n]\frac{d(x\frac{dF}{dx})}{dx}(1,z) = [z^n]\frac{d^2F}{dx}(1,z) + [z^n]\frac{dF}{dx}(1,z) \\ &= [z^n]\frac{2(z^7-6z^6+15z^5-15z^4)}{45(z-1)^3} + E_o(n) \\ &= [z^n]\left(\frac{2}{(z-1)^3}\left(\frac{z^7}{45}-\frac{2z^6}{15}+\frac{z^5}{3}-\frac{z^4}{3}\right)\right) + E_o(n)\,. \end{aligned}$$

If $n > 6$, and using $Var_o(n) = E_{o^2}(n) - E_o^2(n)$, we obtain the variance of o

$$\begin{aligned} \mathrm{V}ar_o(n) &= -\frac{(n-5)(n-6)}{45} + \frac{2(n-4)(n-5)}{15} \\ &\quad -\frac{(n-3)(n-4)}{3} + \frac{(n-2)(n-3)}{3} \\ &\quad +\frac{n}{3} - \frac{n^2}{9} \\ &= \frac{2n}{45}\,. \end{aligned}$$

Note that this is the variance of cherries of independently generated trees. Considering 'linked' trees, i.e. along a recombining chromosome, the variance is smaller.

3.2 The number of pitchforks

The recursive construction presented in Section 3.1.1 can be extended in order to consider also *pitchforks*.

Using different methods, they have been studied before for example by Rosenberg [10]. A pitchfork in a ranked (resp. 0-1-2-increasing) tree is simply a subtree with 3 leaves (resp. 2 nodes). If $r(t)$ denotes the number of pitchforks in $t \in \mathcal{I}_{012}$ the construction of Section 3.1.1 is extended to the new random variable r. We find the following succession rule:

$$\begin{aligned} (o,r,m) &\to (o,r,m+1)^r\,(o,r+1,m+1)^{o-r} \\ (o,r,m) &\to (o+1,r-1,m+1)^r\,(o+1,r,m+1)^{m-2o+1-r}. \end{aligned}$$

Considering now

$$Y(x,v,z) = \sum_{t\in\mathcal{I}_{012}} \frac{x^{o(t)}v^{r(t)}z^{m(t)}}{m(t)\,!},$$

we obtain the following differential equation:

$$(v+x)(v-1)\frac{dY}{dv} = x + xY + x(v-2x)\frac{dY}{dx} + (xz-1)\frac{dY}{dz}. \qquad (8)$$

For $v = 1$ it reduces to eq. (3) but there is non easy analytic solution.

However, we can still obtain the expected value $E_r(m)$ for the number of pitchforks in 0-1-2 increasing trees with m nodes. Starting from (8) and performing the substitutions $x = 1/2$ and $z = 2z$ we obtain

$$\begin{aligned}\frac{dY}{dv}\left(\frac{1}{2}, v, 2z\right) &= \frac{1 + Y\left(\frac{1}{2}, v, 2z\right) + 2(z-1)\frac{dY}{dz}\left(\frac{1}{2}, v, 2z\right)}{2\left(v+\frac{1}{2}\right)(v-1)} \\ &\quad + \frac{\frac{dY}{dx}\left(\frac{1}{2}, v, 2z\right)}{2\left(v+\frac{1}{2}\right)}\end{aligned}$$

from which we have

$$\begin{aligned}[z^m]\frac{dY}{dv}\left(\frac{1}{2}, v, 2z\right) &= [z^m]\frac{Y\left(\frac{1}{2}, v, 2z\right) + 2(z-1)\frac{dY}{dz}\left(\frac{1}{2}, v, 2z\right)}{2\left(v+\frac{1}{2}\right)(v-1)} \\ &\quad + [z^m]\frac{\frac{dY}{dx}\left(\frac{1}{2}, v, 2z\right)}{2\left(v+\frac{1}{2}\right)}.\end{aligned}$$

When $v \to 1$ we find that

$$\begin{aligned}E_r(m) &= [z^m]\left(\lim_{v\to 1}\frac{Y\left(\frac{1}{2}, v, 2z\right) + 2(z-1)\frac{dY}{dz}\left(\frac{1}{2}, v, 2z\right)}{2\left(v+\frac{1}{2}\right)(v-1)}\right) \\ &\quad + \frac{2E_o(m)}{3}.\end{aligned}$$

The considered limit can be determined according to *l' Hospital's rule* taking the derivative of the numerator and the denominator with respect to v and performing then the substitution $v = 1$. Furthermore, from Section 3.1.2 $E_o(m) = (m+1)/3$, and thus

$$
\begin{aligned}
E_r(m) &= [z^m]\left(\frac{1}{3}\sum_{t\in\mathcal{I}_{012}} r(t)\frac{2^{m(t)-o(t)}}{m(t)\,!}z^{m(t)}\right)\\
&\quad +[z^m]\left(\frac{2}{3}(z-1)\sum_{t\in\mathcal{I}_{012}} r(t)m(t)\frac{2^{m(t)-1-o(t)}}{m(t)\,!}z^{m(t)-1}\right)\\
&\quad +\frac{2(m+1)}{9}\\
&= \frac{1}{3}E_r(m)+[z^m]\left(\frac{z-1}{3z}\sum_{k>0}kE_r(k)z^k\right)+\frac{2(m+1)}{9}\\
&= \frac{1}{3}E_r(m)+\frac{mE_r(m)-(m+1)E_r(m+1)}{3}+\frac{2(m+1)}{9}.
\end{aligned}
$$

Reordering terms we obtain the recursion

$$
\begin{aligned}
E_r(2) &= 1;\\
(m+1)E_r(m+1) &= (m-2)E_r(m)+\frac{2(m+1)}{3}.
\end{aligned}
$$

This gives for an increasing tree with $m>2$ nodes

$$
E_r(m)=\frac{m+1}{6}.
$$

From eq. (8) one can also compute the full probability distribution of the random variable r when an increasing tree of fixed size is generated by the coalescent process. Indeed, if we consider

$$
Y_m(x,v,z)=\sum_{t\in\mathcal{I}_{012_m}}\frac{x^{o(t)}v^{r(t)}z^m}{m\,!}
$$

the following result provides a recursion which can be used to compute the functions Y_m for any $m\geq 1$.

Proposition 3 *The following recursion holds:*

$$
\begin{aligned}
Y_1 &= xz\\
Y_{m+1} &= \int\left[(v+x)(1-v)\frac{dY_m}{dv}+xY_m+x(v-2x)\frac{dY_m}{dx}+xz\frac{dY_m}{dz}\right]dz
\end{aligned}
$$

Proof. Consider eq. (8) without the monomial x which appears there. If we then isolate the term $\frac{dY}{dz}$ and integrate both sides of the resulting equation with respect to the variable z we obtain the polynomial Y_{m+1} starting from $Y = Y_m$. □

The results for $m = 1, 2, 3, 4, 5$ are as follows

$$\begin{aligned}
Y_1 &= xz \\
Y_2 &= \frac{1}{2}vxz^2 \\
Y_3 &= \frac{1}{6}vxz^3 + \frac{x^2z^3}{6} \\
Y_4 &= \frac{1}{24}vxz^4 + \frac{x^2z^4}{24} + \frac{1}{8}vx^2z^4 \\
Y_5 &= \frac{1}{120}vxz^5 + \frac{x^2z^5}{120} + \frac{7}{120}vx^2z^5 + \frac{1}{40}v^2x^2z^5 + \frac{x^3z^5}{30}
\end{aligned}$$

The above results concerning cherries and pitchforks can be extended to the joint and conditional probability distributions (see Fig. 4). Summarizing, we state

Proposition 4 *i) The probability of having r' pitchforks in an increasing tree of size m (see Fig. 3) is*

$$P_m(r = r') = [v^{r'}]Y_m\left(\frac{1}{2}, v, 2\right);$$

ii) The probability of having o' cherries and r' pitchforks in an increasing tree of size m is

$$P_m(o = o', r = r') = [x^{o'}v^{r'}]Y_m\left(\frac{x}{2}, v, 2\right);$$

iii) The probability of having r' pitchforks in an increasing tree of size m given it has o' cherries (see Fig. 4) is

$$P_m(r = r'|o = o') = \frac{P_m(o = o', r = r')}{P_m(o = o')} = \frac{[x^{o'}v^{r'}]Y_m\left(\frac{x}{2}, v, 2\right)}{[x^{o'}]Y_m\left(\frac{x}{2}, 1, 2\right)}.$$

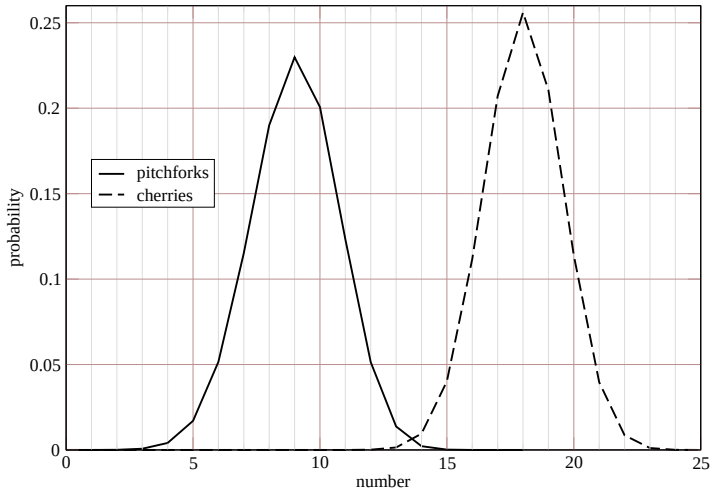


Figure 3: Distributions of cherries and pitchforks for $\mathcal{R}_{54}$ (i.e. $\mathcal{I}_{012_{53}}$).

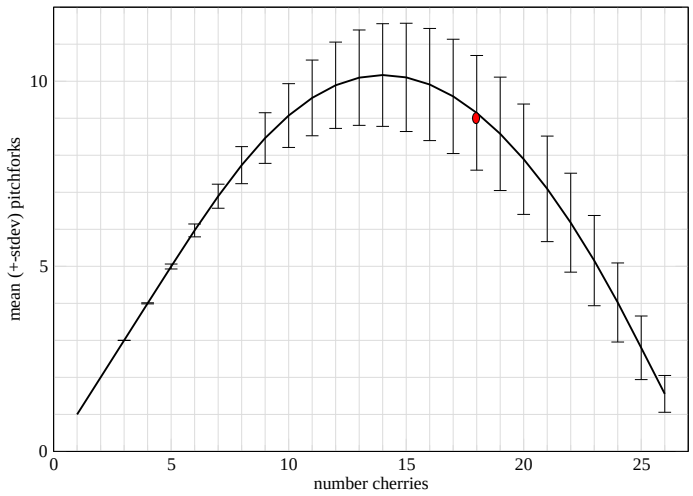


Figure 4: Mean of the conditional probability distribution of pitchforks given the number of cherries for $\mathcal{R}_{54}$.

Acknowledgments

We gratefully acknowledge helpful discussions with L. Ferretti, A. Klassmann and A. Malina. Financial support was provided by the German Research Foundation (DFG-SFB680).

References

[1] C. Banderier, M. Bousquet-Melou, A. Denise, P. Flajolet, D. Gardy, and D. Gouyou-Beauchamps. Generating functions for generating trees. In *Proceedings of 11-th formal power series and algebraic combinatorics*, pages 40–52, 1999.

[2] R. Courant, D. Hilbert. *Methods of Mathematical Physics*. John Wiley & Sons, Inc., 1989.

[3] P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009. URL `http://algo.inria.fr/flajolet/Publications/books.html`.

[4] E. F. Harding. The probabilities of rooted tree-shapes generated by random bifurcation. *Advances in Applied Probability*, 3(1):pp. 44–77, 1971. ISSN 00018678. URL `http://www.jstor.org/stable/1426329`.

[5] Hudson, R. R. (1990). Gene genealogies and the coalescent process. In Oxford Surveys in Evolutionary Biology vol. 7, pp. 1–44. Oxford University Press.

[6] R. R. Hudson. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, 18:337–338, 2002.

[7] J. F. C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13:235–248, 1982.

[8] J. F. C. Kingman. Origins of the coalescent. 1974-1982. *Genetics*, 156 (4):1461–1463, Dec 2000.

[9] A. McKenzie and M. Steel. Distributions of cherries for two models of trees. *Mathematical Biosciences*, 164:81–92, 2000.

[10] N.A. Rosenberg. The mean and the variance of the numbers of r-pronged nodes and r-caterpillars in Yule generated genealogical trees. *Annals of Combinatorics*, 10:129–146, 2006.

[11] N. J. A. Sloane. The on-line encyclopedia of integer sequences. *Notices Amer. Math. Soc.*, 50(8):912–915, 2003. ISSN 0002-9920.

[12] F. Tajima. Evolutionary relationship of DNA sequences in finite populations. *Genetics*, 105(2):437–460, Oct 1983.

[13] J. Wakeley. *Coalescent theory – an introduction*. Roberts&Company, Greenwood Village, Colorado, 2009.